BMC Bioinformatics

BioMed Central

Proceedings

Open Access

Gevab: a prototype genome variation analysis browsing server

Woo-Yeon Kim[†1], Sang-Yoon Kim[†1], Tae-Hyung Kim[†1], Sung-Min Ahn[2,3], Ha Na Byun[1], Deokhoon Kim[2], Dae-Soo Kim[1], Yong Seok Lee[1], Ho Ghang[1], Daeui Park[1], Byoung-Chul Kim[1], Chulhong Kim[1], Sunghoon Lee[1], Seong-Jin Kim*[2] and Jong Bhak*[1]

Addresses: [1]Korean BioInformation Center (KOBIC), KRIBB, Daejeon, Korea, [2]Lee Gil Ya Cancer and Diabetes Institute, Gachon University of Medicine and Science, Incheon, Korea and [3]Department of Translational Medicine, Gachon University Gil Hospital, Incheon, Korea

E-mail: Woo-Yeon Kim - kimplove@kribb.re.kr; Sang-Yoon Kim - sykim@kribb.re.kr; Tae-Hyung Kim - thkim@kribb.re.kr; Sung-Min Ahn - smahn@gachon.ac.kr; Ha Na Byun - tippe@naver.com; Deokhoon Kim - coonya@gmail.com; Dae-Soo Kim - kds2465@kribb.re.kr; Yong Seok Lee - dolsemtl@kribb.re.kr; Ho Ghang - kangho@kribb.re.kr; Daeui Park - daeui@kribb.re.kr; Byoung-Chul Kim - chem1186@kribb.re.kr; Chulhong Kim - chulhong@kribb.re.kr; Sunghoon Lee - ishoon@kribb.re.kr
Seong-Jin Kim* - jasonsjkim@gachon.ac.kr; Jong Bhak* - jongbhak@yahoo.com
*Corresponding author †Equal contributors

from Asia Pacific Bioinformatics Network (APBioNet) Eighth International Conference on Bioinformatics (InCoB2009)
Singapore 7-11 September 2009

Published: 3 December 2009

BMC Bioinformatics 2009, **10**(Suppl 15):S3 doi: 10.1186/1471-2105-10-S15-S3

This article is available from: http://www.biomedcentral.com/1471-2105/10/S15/S3

Abstract

Background: The first Korean individual diploid genome sequence data (KOREF) was publicized in December 2008.

Results: A Korean genome variation analysis and browsing server (Gevab) was constructed as a database and web server for the exploration and downloading of Korean personal genome(s). Information in the Gevab includes SNPs, short indels, and structural variation (SV) and comparison analysis between the NCBI human reference and the Korean genome(s). The user can find information on assembled consensus sequences, sequenced short reads, genetic variations, and relationships between genotype and phenotypes.

Conclusion: This server is openly and publicly available online at http://koreagenome.org/en/ or directly http://gevab.org.

Background

Most known genome browsers, such as NCBI genome [1] and Craig Venter's genome browsers [2], were built for consensus sequences from multiple individuals to construct a reference human genome. Examples of haplotype genome browsers are NCBI, UCSC [3], Ensembl [4], and Venter genome browsers. Recently, the first Asian (Chinese) diploid genome database was published, containing analysis and browsing facilities [5,6]. There are a number of general purpose genome annotation servers. They include Entrez Gene [7], Ensembl genes, OMIM [8] disease associations, HapMap [9], SNPedia

[10], and genetic variations of several individual genomes such as Venter [11], Watson [12], YH (Chinese), and NA18507 (Yoruba) [13]. We have developed an individual genome variation analysis and browsing server (Gevab) for the first Korean personal genome sequence (KOREF).

This server is useful to analyze a diploid human genome produced to study the complex features of human genetic variations. The system integrated multiple variation information such as Venter, Watson, YH, dbSNP, and HapMap genotypes as well as gene information. Hence, users can comparatively study the genotypes in human. Gevab also provides information for SNPs, short indels, and SVs on the KOREF genome. Gevab has two parts: genome variation analysis and genome mapping.

Materials and methods

Data source

KOREF data were generated using the Illumina GA and resulted in 82.73 gigabase (Gb) of sequence (about 1248 million paired 36-base reads and about 504 million 75-base reads).

Using the MAQ (Mapping and Assembly with Qualities) [14] program, these sequences were aligned to the NCBI human genome reference (build 36, without Ns, 2,858,029,377 bp). In total, 99.9% of the NCBI reference genome was covered with an average of 25.92-fold depth (sequencing depth was 28.95-fold).

Database and browser software

In the Gevab Korean genome variation browsing part, the consensus genome sequence and genetic variants include SNPs, short indels, and SVs can be displayed. Gevab used GBrowse [15] developed by GMOD [16] for variation viewing, and the genome map browser part was developed by KOBIC.

Analysis of KOREF

From the KOREF genome sequence, 3.44 millions SNPs were identified and validated using Illumina 1 M-duo and Affy 6.0 BeadChip. We identified 342,965 short indels (-29 - +14 bp). Indels that co-occurred within a window size of 20 bp were filtered out, since they were primarily from length polymorphisms in homopolymeric tracts of A or T. Using paired-end reads, we found 2920 deletions and 415 inversion structural variants (SV) in the range of 0.1~100 kb. In addition, we detected 963 insertion events in the range of 175~250 bp. These insertions are present in the KOREF genome but absent in the NCBI reference genome. MySql and PHP, python, and AJAX were used in database construction and interface utility.

Results

Features of Gevab

The Gevab has genome variation analysis and genome map browser parts. The genome variation analysis part contains external public data sources, including the reference sequence of the human genome ((NCBI build 36), the Ensembl gene annotation, the Entrez gene annotations, dbSNP ver. 129 [17], OMIM annotations, and SNP frequencies of the HapMap population as well as genotype, indel, and structure variation of the KOREF. It is also integrated with other individual SNP variants such as James Watson's, Craig Venter's, and YangHuang's genotypes (Table 1). These external data sets are coordinated with the NCBI reference genome. A search can be done by putting in a genome location, a gene symbol, a RefSeq id, a dbSNP id, or an Ensembl gene id. When the user searches Gevab with a query, a graphical view of a chromosome ideogram and contigs are displayed. The gene locations within the 2 MB region centered on the query are also represented. For the displayed region in our browser, users can also download data with gff or fasta format ftp://ftp.kobic.re.kr/pub/KOBIC-KoreanGenome/.

Table 1: Features of Gevab, Venter, Watson, and YH genome browsers. Availability of features is indicated by "O" for "yes" and "X" for "no."

	Gevab	Venter	Watson	YH
genome	Korean	Caucasian	Caucasian	Chinese
read mapping	O	X	X	O
sequencing coverage	O	O	O	X
genotype	O	O	O	O
indel	O	O	X	X
structure variation	O	O	X	X
variations to compare	Venter, Watson, YH, dbSNP, HapMap	dbSNP	dbSNP, HapMap	dbSNP, HapMap
web site	http://gevab.org	http://huref.jcvi.org	http://jimwatsonsequence.cshl.edu	http://yh.genomics.org.cn/

Gevab's map browser

The genome map browser provides reads mapping and quality information obtained from a personal genome project. A search can be done by chromosomal position. The width of a displayed region can be controlled. The browser also has zoom in and out and left and right movement functions.

When a user a chooses 1000 bp window size or longer, the browser shows a graphical view with forward and reverse pair-end reads in dark and light green, and single reads in red (Figure 1(A)). For shorter than 1000 bp window size, the browser is converted to a text mode that additionally shows quality information of mapped reads (Figure 1(B)).

Gevab's variation browser

To study genome variations, a genome variation browser is more useful than a genome map browser. As an example, if a user is interested in the "NOC2L" gene, s/he can get KOREF, Watson, YH, and Venter genome variation information through the variation browser part (Figure 2).

KOREF data access

The KOREF database is developed and maintained by KOBIC (Korean Bioinformation Center). The database contains all the raw and processed data of KOREF, including KOREF consensus sequence, genetic variants, and short read alignments. These data are available for downloading. The KOREF data have been deposited in

Figure 1
A screenshot of the genome map browser. (A) Graphic mode (B) Text mode.

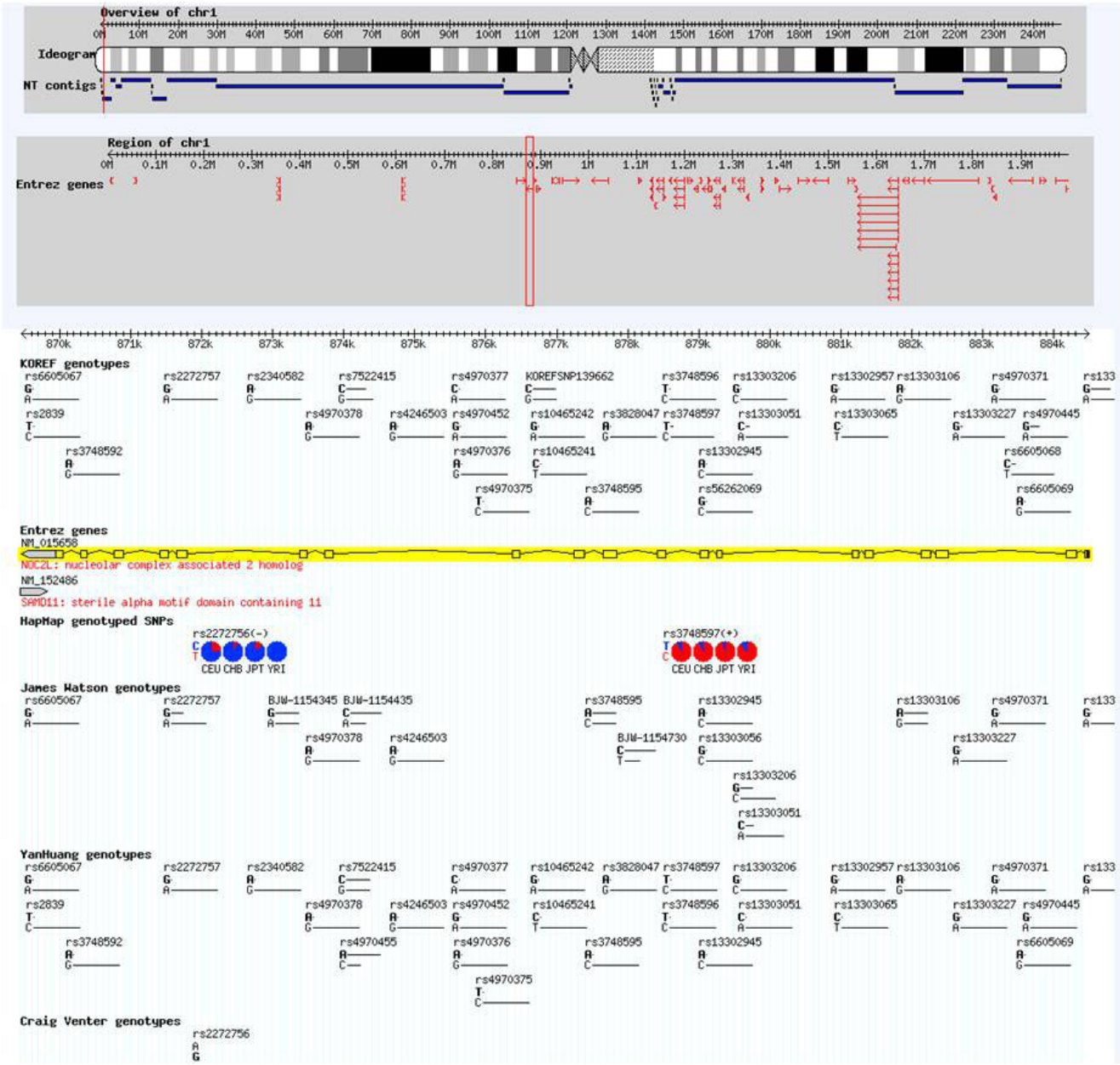

Figure 2
A screenshot of the genome variation browser.

the NCBI Short Read Archive (Accession Number SRA008175).

Conclusion

Gevab contains all the raw and processed data of a Korean genome sequence, variants, and annotation. Gevab provides open and public access to all data of an individual personal diploid genome.

The variation browser part was designed to present genetic variant evidence, including the position, number, and status of reads, GC content, and several mapping information. These provide valuable detailed information such as comparison and validation of genetic variations to further communities for sequencing individual genome.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

WYK developed the genetic variation browser and helped to write manuscript. SYK developed the genome

map browser. THK wrote manuscript. SMA and DK provided the KOREF data. HNB, DSK, YSL, HG, DP, BCK, CK, and SL provided counseling on issues related to Gevab development. SJK supervised the whole project and guided to production of KOREF data. JB supervised the bioinformatic analysis and manuscript writing.

Note

Other papers from the meeting have been published as part of *BMC Genomics* Volume 10 Supplement 3, 2009: Eighth International Conference on Bioinformatics (InCoB2009): Computational Biology, available online at http://www.biomedcentral.com/1471-2164/10?issue=S3.

Acknowledgements

This work was supported by a grant from the KRIBB Research Initiative Program of Korea, by the Korea Science and Engineering Foundation (KOSEF) grant funded by the Korean government (MOST), the National Research Foundation of Korea (NRF) grant (No. R11-2008-044-03004-0, S.M.A.), a grant from Ministry of Knowledge Economy (Standard Reference Data Program), and generous funding from the Gachon University of Medicine and Science & Gachon University Gil Hospital. We thank Ryu Gichan for crucial administration assistance, Ryu Jeawoon and Cho Suan for web application, and Maryana Bhak for editing.

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 15, 2009: Eighth International Conference on Bioinformatics (InCoB2009): Bioinformatics. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/10?issue=S15.

References

1. **Mapviewer.** http://www.ncbi.nlm.nih.gov/projects/mapview/.
2. Axelrod N, Lin Y, Ng PC, Stockwell TB, Crabtree J, Huang J, Kirkness E, Strausberg RL, Frazier ME and Venter JC, *et al*: **The HuRef Browser: a web resource for individual human genomics.** *Nucleic Acids Res* 2009, **37 Database:** D1018–1024.
3. Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A and Pheasant M, *et al*: **The UCSC Genome Browser Database: update 2009.** *Nucleic Acids Res* 2009, **37 Database:** D755–761.
4. Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P and Clarke L, *et al*: **Ensembl 2009.** *Nucleic Acids Res* 2009, **37 Database:** D690–697.
5. Wang J, Wang W, Li R, Li Y, Tian G, Goodman L, Fan W, Zhang J, Li J and Guo Y, *et al*: **The diploid genome sequence of an Asian individual.** *Nature* 2008, **456(7218):**60–65.
6. **Chinese genome browser.** http://yh.genomics.org.cn/.
7. Maglott D, Ostell J, Pruitt KD and Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2007, **35 Database:** D26–31.
8. Amberger J, Bocchini CA, Scott AF and Hamosh A: **McKusick's Online Mendelian Inheritance in Man (OMIM).** *Nucleic Acids Res* 2009, **37 Database:** D793–796.
9. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P and Leal SM, *et al*: **A second generation human haplotype map of over 3.1 million SNPs.** *Nature* 2007, **449(7164):**851–861.
10. **SNPedia.** http://www.snpedia.com/.
11. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, Walenz BP, Axelrod N, Huang J, Kirkness EF and Denisov G, *et al*: **The diploid genome sequence of an individual human.** *PLoS Biol* 2007, **5(10):**e254.
12. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V and Roth GT, *et al*: **The complete genome of an individual by massively parallel DNA sequencing.** *Nature* 2008, **452(7189):**872–876.
13. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL and Bignell HR, *et al*: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456(7218):**53–59.
14. Li H, Ruan J and Durbin R: **Mapping short DNA sequencing reads and calling variants using mapping quality scores.** *Genome Res* 2008, **18(11):**1851–1858.
15. Stein LD, Mungall C, Shu S, Caudy M, Mangone M, Day A, Nickerson E, Stajich JE, Harris TW and Arva A, *et al*: **The generic genome browser: a building block for a model organism system database.** *Genome Res* 2002, **12(10):**1599–1610.
16. **GMOD.** http://gmod.org.
17. Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM and Sirotkin K: **dbSNP: the NCBI database of genetic variation.** *Nucleic Acids Res* 2001, **29(1):**308–311.